Los Alamos
National Laboratory

Fat-shattering of affine functions

Don Hush and Clint Scovel
Modeling, Algorithms and Informatics Group, CCS-3
Mail Stop B265
Los Alamos National Laboratory
Los Alamos, NM 87545
(dhush,jcs)@lanl.gov

LANL Technical Report: LA–UR–03–0937

Report Date: February 13, 2003

Abstract

We compute the fat-shattering function and the level fat-shattering function for important classes of affine functions. We observe that the level fat-shattering function and the fat-shattering function are identical for these classes. In addition we observe that the notion that adding the constant term to linear functions increases the dimension by at most 1 is incorrect for fat-shattering and level fat-shattering.

Key Words

fat shattering, level fat shattering, affine functions

Fat-shattering was introduced by Kearns et al. (Kearns & Schapire, 1994) to provide lower bounds on sample complexity for a learning problem. However it appears that fat shattering is useful more generally. In particular Alon et. al. (Alon, Ben-David, Cesa-Bianchi, & Haussler, 1997) have proven a generalization of Sauer's lemma, bounding covering numbers of a function class in terms of its fat-shattering dimension, leading to a new characterization of Glivenko Cantelli classes. Shawe-Taylor et al. (Shawe-Taylor, Bartlett, Williamson, & Anthony, 1998) use the lemma of Alon et al. to provide the first justification for the performance of margin based classifiers such as Vapnik's support vector machines (Vapnik, 1998). Consequently it appears useful to understand the fat-shattering dimension of function classes.

Shawe-Taylor et al. (Shawe-Taylor et al., 1998) show that when B_R is the ball of radius R in a pre-Hilbert space H (i.e. there is no completeness requirement) of dimension n and $\mathcal{F}$ is the class of affine functions $f(x) = \phi \cdot x + c$ with $|\phi|_H \leq 1$ and $|c| \leq R$ then

$$fat_{\mathcal{F}}(\gamma) \leq \min\left(\frac{9R^2}{\gamma^2}, n+1\right) + 1.$$

In this paper we eliminate the constraint on c and compute the shattering function *exactly.* As a corollary we obtain a simple approximate form for the shattering function which sharpens this bound essentially to an equality

$$\max\left(\frac{R^2}{\gamma^2}, 1\right) \leq fat_{\mathcal{F}}(\gamma) \leq \min\left(\frac{R^2}{\gamma^2} + \frac{5}{4}, n+1\right).$$

The proof relies on that used in Hush and Scovel (Hush & Scovel, 2001) to prove a theorem of Vapnik on the *level* fat-shattering dimension of the affine functions. It is modeled on that found in Bartlett et al. (Bartlett & Shawe-Taylor, 1999) which they applied to the linear case. Their technique seems to be based on Gurvits (Gurvits, 1997). In the process of proof we obtain that the fat shattering function and the *level* fat shattering function are identical for the affine functions. Comparing with bounds of Bartlett et. al(Bartlett & Shawe-Taylor, 1999) for linear functions we observe that the notion that going from linear to affine functions should add at most 1 to the fat-shattering or the level fat-shattering dimension is incorrect.

Definition 1. Let $\mathcal{F}$ denote a set of real valued functions on a set X and consider $\gamma > 0$. We say that a subset $A \subset X$ is γ-shattered by $\mathcal{F}$ if there is a real vector parameterized by A, $\{r_a \in \mathbb{R} : a \in A\}$, such that for all binary vectors b parameterized by A, there is a function f_b such that $f_b(a) \geq r_a + \gamma$ when $b_a = 1$ and $f_b(a) \leq r_a - \gamma$ otherwise. The fat-shattering dimension of $\mathcal{F}$ at scale γ, $fat_{\mathcal{F}}(\gamma)$, is the size of the largest subset $A \subset X$ which is γ-shattered by $\mathcal{F}$.

The definition of γ *level* shattering is the same as above but where $r_a = r$ is constant in a.

We now state and prove our main result.

Theorem 1. *Let H denote a prehilbert space of dimension n and consider its closed ball $X = B_{R_1} \subset H$ of radius R_1. Let $\mathcal{F}$ denote the class of functions on X defined by $f(x) = \phi \cdot x + c$ with $|\phi|_H \leq R_2$ and $c \in \mathbb{R}$. Denote*

$$\gamma_k = \frac{1}{\sqrt{k-1}} \quad k \ \ even$$

$$\gamma_k = \frac{k}{k-1}\frac{1}{\sqrt{k+1}} \quad k \;\; odd \; .$$

Then

$$fat_{\mathcal{F}}(\gamma) = n+1, \quad \frac{\gamma}{R_1 R_2} \leq \gamma_{n+1} \tag{1}$$

$$fat_{\mathcal{F}}(\gamma) = k, \quad \gamma_{k+1} < \frac{\gamma}{R_1 R_2} \leq \gamma_k, \quad 1 \leq k \leq n+1 \; . \tag{2}$$

This function is correct when $n = \infty$ but since $\gamma > 0$ the first line (1) is then void.

Proof. We prove the theorem for $R_1 = R_2 = 1$. The general theorem then follows from simple scaling arguments. We first prove that

$$1 \leq fat_{\mathcal{F}}(\gamma) \leq n+1. \tag{3}$$

The inequality $1 \leq fat_{\mathcal{F}}(\gamma)$ follows from the fact that c can be any real number. We now assume n is finite for otherwise there is nothing to prove. Let $A = \{x_1, .., x_k\}$ where $x_i \in X, i = 1, .., k$. From the definition, if A is γ-shattered by $\mathcal{F}$ then there exists k real values $r_i, i = 1, .., k$ such that for every binary vector $b = (b_1, .., b_k) \in \{-1, 1\}^k$ there is a (ϕ_b, c_b) with $|\phi_b| \leq 1$ such that

$$\phi_b \cdot x_i + c_b \geq r_i + \gamma, \quad b_i = 1 \tag{4}$$

$$\phi_b \cdot x_i + c_b \leq r_i - \gamma, \quad b_i = -1 \; . \tag{5}$$

If we define $\tilde{x} = (x, r)$, $\tilde{x}_i = (x_i, r_i)$, $\tilde{\phi} = (\phi, -1)$ and $\tilde{\phi}_b = (\phi_b, -1)$ then we obtain

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \geq \gamma, \quad b_i = 1$$

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \leq -\gamma, \quad b_i = -1 \; .$$

Since $\gamma > 0$ this implies that

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b \geq 0, \quad b_i = 1 \tag{6}$$

$$\tilde{\phi}_b \cdot \tilde{x}_i + c_b < 0, \quad b_i = -1 \; . \tag{7}$$

The set of functions $\tilde{x} \mapsto \tilde{\phi} \cdot \tilde{x} + c = \phi \cdot x - r + c$ as (ϕ, c) vary is a subset of an affine space of dimension $\leq n+1$. Applying the following generalization of a theorem of Steele and Dudley, (see Theorem 13.9 of (Devroye, Györfi, & Lugosi, 1996)) to this affine space of functions satisfying (6) and (7) finishes the proof of (3). We note that this proof and therefore the corresponding bound also applies for level fat-shattering.

Lemma 1. *Let $\mathcal{G}$ be an affine space of real functions on a set X with $dim(\mathcal{G}) = r$. The class of sets*

$$\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$$

has VC dimension $V_{\mathcal{A}} \leq r$.

Proof. The proof follows that of Theorem 13.9 in (Devroye et al., 1996). It suffices to show that no set of size $k > r$ points can be shattered by $\mathcal{A}$. Fix arbitrary points $x_1, .., x_k$ in X and consider the linear mapping L defined by

$$g \mapsto (g(x_1), .., g(x_k)) \in \mathbb{R}^k.$$

Since $\mathcal{G}$ is affine and L is linear $L\mathcal{G}$ is an affine subspace of $\mathbb{R}^k$ with $dim(L\mathcal{G}) \leq r < k$. If this affine subspace contains the origin, then there is a nonzero vector α orthogonal to $L\mathcal{G}$ with at least one positive value. If we define $S = \{i : \alpha_i \leq 0\}$ then the set of points $\{x_i : \alpha_i \leq 0\}$ can never equal a set $\{x_i : g(x_i) \geq 0\}$ for any g because the right hand side of

$$0 =< \alpha, Lg >= \sum_i \alpha_i g(x_i) = \sum_S \alpha_i g(x_i) + \sum_{S^c} \alpha_i g(x_i)$$

would be strictly negative giving a contradiction. Consequently $\mathcal{A}$ cannot shatter $x_1, .., x_k$. Similarly, if the affine subspace $L\mathcal{G}$ does not contain the origin a vector α in $L\mathcal{G}$ closest to the origin satisfies

$$< \alpha, (Lg - \alpha) >= 0, \quad \forall g \in \mathcal{G}$$

which means that

$$< \alpha, Lg >= \sum_i \alpha_i g(x_i) = |\alpha|^2 > 0.$$

If we define $S = \{i : \alpha_i \leq 0\}$ then the set of points $\{x_i : \alpha_i \leq 0\}$ can never equal a set $\{x_i : g(x_i) \geq 0\}$ for any g because the left hand side of

$$\sum_i \alpha_i g(x_i) = \sum_S \alpha_i g(x_i) + \sum_{S^c} \alpha_i g(x_i) > 0$$

would be strictly negative giving a contradiction. Consequently $\mathcal{A}$ cannot shatter $x_1, .., x_k$.

♦

We now proceed to obtain stronger inequalities.

Lemma 2. *Suppose that $k > 1$ points are contained in the ball of radius 1 in H and are γ fat-shattered by the affine linear functions $\phi \cdot x + c$ where $|\phi| \leq 1$ and $c \in \mathbb{R}$. Then*

$$\gamma \leq \gamma_k \; .$$

We note that Hush and Scovel (Hush & Scovel, 2001) prove the same inequalities of Lemma 2 for level fat-shattering.

Proof. Let the binary vector b determine a nontrivial partition of the index set. We define the weight

$$L(b)_i = \frac{1}{|\{j : b_j = b_i\}|}$$

so that

$$\sum_{i=1}^{k} L(b)_i = 2$$

and

$$\sum_{i=1}^{k} L(b)_i b_i = 0.$$

If the k points x_i are γ shattered then there is a k-vector r such that for each binary vector b there is a (ϕ_b, c_b) that satisfies (4) and (5). These two equations can be written

$$b_i \phi_b \cdot x_i + b_i c_b \geq b_i r_i + \gamma \quad 1 = 1, .., k. \tag{8}$$

Since $L(-b) = L(b)$, the sum $\sum L(-b)_i(-b)_i r_i = -\sum L(b)_i b_i r_i$ is odd with respect to reflection and among b and $-b$ at least one satisfies

$$\sum L(b)_i b_i r_i \geq 0.$$

Consider this choice of b. Since the weights $L(b)_i$ are positive, we multiply Equation 8 by $L(b)_i$ and sum to obtain

$$\phi_b \cdot \sum L(b)_i b_i x_i \geq 2\gamma$$

where we have utilized the fact that $\sum_{i=1}^{k} L(b)_i = 2$ and $\sum_{i=1}^{k} L(b)_i b_i = 0$. Since $|\phi_b| \leq 1$ the Cauchy-Schwartz inequality implies that

$$|\sum L(b)_i b_i x_i| \geq 2\gamma.$$

Since $\sum L(-b)_i(-b)_i x_i = -\sum L(b)_i b_i x_i$ this inequality is also true for $-b$ and consequently for all b representing non trivial partitions.

Hush and Scovel (Hush & Scovel, 2001) show that when k is even

$$E(|\sum L(b)_i(b)_i x_i|^2) \leq \frac{4}{k-1}$$

where E denotes expectation with respect to the uniform distribution over all partitions of the k indices into two $\frac{k}{2}$ sized subsets. They also showed that when k is odd,

$$E(|\sum L(b)_i(b)_i x_i|^2) \leq \frac{4k^2}{(k-1)^2(k+1)}$$

where E denotes expectation with respect to the uniform distribution over all partitions into a size r subset and a size $r+1$ subset where $r = \frac{k+1}{2}$.

When k is even there must exist a partition b such that

$$|\sum L(b)_i(b)_i x_i|^2 \leq E(|\sum L(b)_i(b)_i x_i|^2) \leq \frac{4}{k-1}$$

but since we know that

$$|\sum L(b)_i b_i x_i| \geq 2\gamma$$

we obtain that

$$\gamma \leq \frac{1}{\sqrt{k-1}} \quad k \ \ even.$$

Likewise, when k is odd, there must exist a partition such that

$$|\sum L(b)_i (b)_i x_i|^2 \leq E(|\sum L(b)_i (b)_i x_i|^2) \leq \frac{4k^2}{(k-1)^2(k+1)}$$

giving the inequality

$$\gamma \leq \frac{k}{k-1}\frac{1}{\sqrt{k+1}} \quad k \ \ odd.$$

The proof of Lemma 2 is finished.

♦

We now complete the proof of Theorem 1. Since the inequalities of Lemma 2 and the bound (3) apply for level fat-shattering it is not hard to see that the following argument constructs both the fat-shattering function and the level fat-shattering function simultaneously and that they are identical. Vapnik (Vapnik, 1998) observed that when $k \leq n+1$ is finite the regular unit k-simplex achieves

$$\gamma = \gamma_k$$

so that $fat_{\mathcal{F}}(\gamma_k) \geq k$. Since $fat_{\mathcal{F}}$ is monotonic in γ we obtain that

$$\gamma \leq \gamma_k \Rightarrow fat_{\mathcal{F}}(\gamma) \geq k, \quad k \leq n+1. \tag{9}$$

The inequalities of Lemma 2 imply that

$$\gamma > \gamma_k \Rightarrow fat_{\mathcal{F}}(\gamma) < k, \quad 1 < k\ . \tag{10}$$

Combining (9) and (10) we obtain

$$\gamma_{k+1} < \gamma \leq \gamma_k, \quad 1 < k \leq n+1 \ \Rightarrow fat_{\mathcal{F}}(\gamma) = k \tag{11}$$

and from (9), the monotonicity of $fat_{\mathcal{F}}$, and the bound (3) we obtain

$$\gamma \leq \gamma_{n+1} \Rightarrow fat_{\mathcal{F}}(\gamma) = n+1$$

and so obtain Equations (1) and (2) restricted to $k > 1$. It follows from (11), the monotonicity of $fat_{\mathcal{F}}$, and the lower bound of (3) that $fat_{\mathcal{F}}(\gamma) = 1$ when $\gamma > 1$. The proof of Theorem 1 is finished. ♦

We now proceed to make the function (1,2) of Theorem 1 explicit.

Corollary 1. *Let H denote a prehilbert space of dimension n and consider its closed ball $X = B_{R_1} \subset H$ of radius R_1. Let $\mathcal{F}$ denote the class of functions on X defined by $f(x) = \phi \cdot x + c$ with $|\phi|_H \leq R_2$ and $c \in \mathbb{R}$. Then*

$$\max\left(\frac{R_1^2 R_2^2}{\gamma^2}, 1\right) \leq fat_{\mathcal{F}}(\gamma) \leq \min\left(\frac{R_1^2 R_2^2}{\gamma^2} + \frac{5}{4}, n+1\right).$$

Proof. Again we prove the corollary for $R_1 = R_2 = 1$ and the general theorem then follows from simple scaling arguments.

Equation (2) becomes

$$fat_{\mathcal{F}}(\gamma) = k, \quad \frac{k+1}{k}\frac{1}{\sqrt{k+2}} < \gamma \leq \frac{1}{\sqrt{k-1}}, \quad k \text{ even} \tag{12}$$

$$fat_{\mathcal{F}}(\gamma) = k, \quad \frac{1}{\sqrt{k}} < \gamma \leq \frac{k}{k-1}\frac{1}{\sqrt{k+1}}, \quad k \text{ odd} \tag{13}$$

To finish the proof of Corollary 1 we establish

$$fat_{\mathcal{F}}(\gamma) - \frac{1}{\gamma^2} \in (0, \frac{5}{4}] \,. \tag{14}$$

To that end, consider the case where $fat_{\mathcal{F}}(\gamma) = k$ is even. We prove the sharper statement

$$fat_{\mathcal{F}}(\gamma) - \frac{1}{\gamma^2} \in (\frac{n+1}{(n+2)^2}, 1], \quad fat_{\mathcal{F}}(\gamma) \text{ even} \,. \tag{15}$$

Equation (12) is equivalent to

$$\frac{k}{(k+1)^2} < k - \frac{1}{\gamma^2} \leq 1$$

Since $fat_{\mathcal{F}}(\gamma) \leq n+1$ and $\frac{k}{(k+1)^2}$ is monotonically decreasing we obtain the claim (15) for finite n. It is not hard to see that the result is also correct when $n = \infty$. Now consider the case where $fat_{\mathcal{F}}(\gamma) = k$ is odd. Equation (13) is equivalent to

$$0 < k - \frac{1}{\gamma^2} \leq 1 + \frac{1}{k} - \frac{1}{k^2}. \tag{16}$$

Since the function $1 + \frac{1}{k} - \frac{1}{k^2}$ has a maximum value $\frac{5}{4}$ we conclude

$$fat_{\mathcal{F}}(\gamma) \in \frac{1}{\gamma^2} + (0, \frac{5}{4}], \quad fat_{\mathcal{F}}(\gamma) \text{ odd} \tag{17}$$

Note that inequality (16) is sharp in the sense that the choice $\gamma = \frac{k}{k-1}\frac{1}{\sqrt{k+1}}$ achieves the upper bound and since the function $1 + \frac{1}{k} - \frac{1}{k^2}$ is strictly greater than 1 for $k > 1$ we conclude that the often stated bound

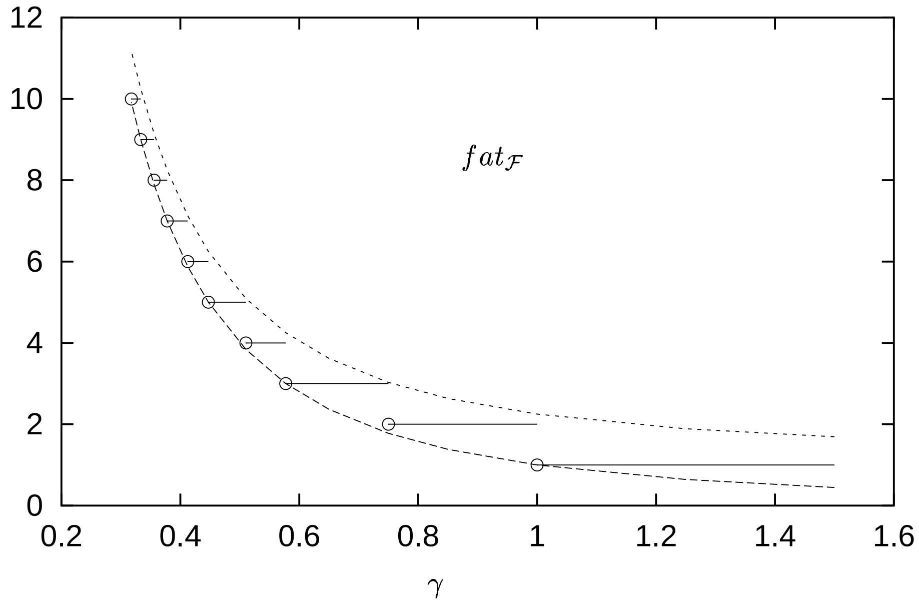
$$fat_{\mathcal{F}}(\gamma) \leq \frac{1}{\gamma^2} + 1 \tag{18}$$

is incorrect.

Combining the even (15) and the odd (17) cases implies the claim (14) and together with (3) completes the proof of Corollary 1.

♦

Below we present a graph of the function $fat_{\mathcal{F}}$ along with the upper and lower bounds implied by Corollary 1.



References

Alon, N., Ben-David, S., Cesa-Bianchi, N., & Haussler, D. (1997). Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM*, *44*(4), 615–631.

Bartlett, P. L., & Shawe-Taylor, J. (1999). Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: Support vector learning.* MIT Press.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition.* New York, NY: Springer.

Gurvits, L. (1997). A note on the scale sensitive dimension of linear bounded functionals in banach spaces. *Proceedings of Algorithm Learning Theory, ALT-97.*

Hush, D., & Scovel, C. (2001). On the vc dimension of bounded margin classifiers. *Machine Learning*, *45*, 33–44.

Kearns, M., & Schapire, R. E. (1994). Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, *48*(3), 464–497.

Shawe-Taylor, J., Bartlett, P., Williamson, R. C., & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, *44*(5), 1926–1940.

Vapnik, V. N. (1998). *Statistical learning theory.* New York: John Wiley and Sons, Inc.